Semantics and Complexity of Question Answering Systems: Towards a Moore's Law for Natural Language Engineering

Amit Bagga
General Electric CRD,
PO Box 8
Schenectady, NY 12301
bagga@crd.ge.com

Wlodek Zadrozny
IBM Research,
T. J. Watson Research Center
Yorktown Heights, NY 10598
wlodz@us.ibm.com

James Pustejovsky
Department of Computer Science
Volen Center for Complex Systems
Brandeis University
Waltham, MA 02254
jamesp@cs.brandeis.edu

Panel Description

Motivation:

1. There's a proliferation of QA systems and NL chat systems on the Web and in intranets (e.g. www.neuromedia.com or www.ask.com). The question arises: how far are these systems from "real" NLU? After all, they seem to follow Eliza. Do they? How do we settle such questions?
2. QA systems typically comprise of an NL understanding systems and a knowledge base/database. Given a user's query, what is the complexity of retrieving the information from the database? How do we approach this problem? What does it mean that one query is more complex than the other?
3. NLP is in demand. For this demand to be sustained we need a *"Moore's Law for natural language engineering"*, so that the business community could understand what is possible when and at what price.

This panel will address the above three issues from three perspectives:
a. lexicon, in particular complexity of lexical semantics.
b. text and dialog, and their semantics complexity
c. abstract model, which supports the analyses presented in (a) and (b).

Regarding the first question, we want to address the issue of systems evaluation in a systematic way: starting with the lexicon, we can define general metrics for entity and relation. On the level of textual databases or collection of facts, we can compare their structure by using the concepts of semantic complexity of text genre and corpora. On the level of accessing this information, we need both a theoretical foundation and an empirical study of the task difficulty. The theory we propose postulates a NL semantics based on the idea that the meaning of text (or corpus) is defined with respect to an automaton which answers a class of questions about it (e.g. what is X? how do I do Y?). The empirical study shows it is possible to classify facts (information) into levels each signifying different degrees of difficulty of extracting facts from text. The study also shows that there is a great variation in the ways a fact can be expressed (in different texts). In fact, the study, by analyzing the output of Information Extraction systems, verifies a very intuitive hypothesis that there are facts which are simply harder to extract than others. From the perspective of the three panelists, the complexity of a QA system in which users retrieve data from a database should be analyzed on several levels: A. the level of complexity of background knowledge needed to understand the content of the query and the database entry: "what"-questions can be more easily answered than "why-"questions, because the latter require not only static background knowledge, but also the knowledge of plans and goals; B. the complexity of the vocabulary and syntax of the text: "what-"questions can be more easily answered if they appear in simpler syntactic constructions. C. the complexity of interaction: e.g. it is easier to answer one type of questions than to manage a conversation in which one question can lead to another, contextually dependent question.

Question 3 is another way of phrasing Question 2. It brings the business perspective to the issue. While the panelists do not formulate a *"Moore's Law for Natural Language Engineering"*, they hope that this panel can provide an impetus and some tools to research the issue.

Theses and questions for a panel discussion and discussion with the audience:

- NL engineering is in the need of metrics: for system complexity, task difficulty, estimated effort, etc.
  - Can software metrics and methodologies apply?
  - How do we take into account background knowledge?
  - How do we go about formulating a Moore's Law for NLE?
- NL systems, especially QA systems, should be positioned as interaction systems in P.Wegners sense:
  - NL systems operate on streams,

- NL grammars can be viewed as collections of objects (in the software engineering sense).

- NL semantics should be contextual and task dependent
  - It is possible to define a semantics based on interactive, task dependent QA behavior.
  - Such a semantics is consistent with the compositionality postulate
- NL semantics should be modular and task independent; such a semantics can support different tasks with minimum customization.
- Can these two views coexist?